A Keyframe-Based Approach for Auditing Bias in YouTube Shorts Recommendations

Mert Can Cakmak
Computer and Information Science
University of Arkansas - Little Rock
Little Rock, Arkansas, USA
mccakmak@ualr.edu

Nitin Agarwal
ICSI, University of California, Berkeley
Berkeley, California, USA
COSMOS Research Center
University of Arkansas - Little Rock
Little Rock, Arkansas, USA
nxagarwal@ualr.edu

Abstract

YouTube Shorts and other short-form video platforms now influence how billions engage with content, yet their recommendation systems remain largely opaque. Small shifts in promoted content can significantly impact user exposure, especially for politically sensitive topics. In this work, we propose a keyframe-based method to audit bias and drift in short-form video recommendations. Rather than analyzing full videos or relying on metadata, we extract perceptually salient keyframes, generate captions, and embed both into a shared content space. Using visual mapping across recommendation chains, we observe consistent shifts and clustering patterns that indicate topic drift and potential filtering. Comparing politically sensitive topics with general YouTube categories, we find notable differences in recommendation behavior. Our findings show that keyframes provide an efficient and interpretable lens for understanding bias in short-form video algorithms.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

YouTube Shorts, recommender systems, algorithmic bias, content drift, keyframe analysis, content auditing

1 Introduction

Short-form video platforms like YouTube Shorts have rapidly reshaped online content consumption, offering users a continuous, swipe-driven stream of personalized video recommendations. Since its global launch in 2021, YouTube Shorts has grown to over two billion monthly users, making it one of the most influential content delivery systems worldwide. As algorithms decide what users see next, even small recommendation biases can significantly affect exposure to underrepresented or sensitive topics.

Prior studies have highlighted ideological skew, homogenization, and popularity bias in recommender systems, often through metadata or user interaction data. However, these approaches overlook the actual visual content users consume, and analyzing entire videos at scale is computationally intensive. This challenge is especially pronounced in short-form ecosystems where content changes rapidly.

To address this, we introduce a keyframe-based method for analyzing recommendation behavior in YouTube Shorts. Using the PRISM framework, we extract perceptually important keyframes from both seed and recommended videos. We then use CLIP to embed these frames, along with captions generated using the Llama-3.2-11B-Vision-Instruct model, and project them into a shared space using UMAP. This allows us to observe how recommendations visually and thematically shift from the original content.

We collect data across two domains: videos related to the politically sensitive "South China Sea" (SCS) and a control group sampled from general YouTube video categories. The SCS topic is frequently subject to geopolitical framing and content moderation, making it a meaningful case for studying potential recommendation bias. In contrast, general content provides a baseline for understanding typical recommendation behavior. This contrast allows us to assess whether algorithmic dynamics differ when political sensitivity is involved.

Our study is guided by the following research questions:

- **RQ1:** Do YouTube Shorts recommendations visually or thematically shift away from their seed content?
- **RQ2:** Can keyframes and their captions reveal patterns in recommended content that suggest potential bias or filtering?
- **RQ3:** Does the degree of content shift differ between politically sensitive topics and general categories?

Our findings show that keyframes provide a focused and interpretable lens for studying short-form recommendations. We observe measurable divergence and clustering, with more pronounced shifts in sensitive-topic cases. This work contributes a scalable, content-driven method for auditing algorithmic behavior in modern video platforms.

2 Literature Review

YouTube's recommendation system, powered by deep neural networks, optimizes for predicted watch time but has raised concerns about various biases and drifts [11]. These include selection, popularity, and position bias, as well as evolving issues like recommendation drift, algorithmic drift, and content drift [16, 8, 14]. Empirical audits reveal ideological and emotional skews, right-leaning accounts receive more radical content, while emotionally negative or morally charged videos are filtered out deeper in the chain [18, 25, 17]. For sensitive topics such as China–Uyghur, recommendations increasingly favor positive sentiment content, suppressing moral critique [4, 6]. Network analyses show that recommendation structures reinforce thematic silos, limiting content diversity [21, 22, 23].

Short-form platforms like YouTube Shorts and TikTok intensify these issues. Studies find rapid content drift from controversial seeds to apolitical or visually neutral recommendations [3]. TikTok audits reveal partisan bias favoring Republican content [19], while Kuaishou exhibits duration bias favoring longer videos [39, 37, 28]. Algorithmic feedback loops further reinforce mainstream narratives, reducing minority exposure [29]. Mitigation strategies include re-ranking, causal modeling, and adversarial learning to address ideological, temporal, and duration-based biases in recommendation pipelines [18, 37, 39]. However, most existing efforts rely on full video analysis or engagement traces, with limited focus on visual-semantic dynamics in short-form formats like YouTube Shorts.

We present the first keyframe-based approach of bias and drift in YouTube Shorts. Unlike prior work that processes full videos or relies solely on metadata, we extract perceptually salient keyframes as concise visual summaries. These serve as effective proxies for comparing content across recommendation chains, revealing shifts in visual framing and content representation. This approach enables a faster and more focused analysis of recommendation behavior in short-form video systems.

3 Data Collection

To examine potential algorithmic and recommendation biases in YouTube Shorts, we collected data across two domains: (1) the South China Sea territorial disputes, and (2) general content categories reflecting YouTube's broader ecosystem. This dual focus enables comparison between politically sensitive and neutral domains, allowing us to assess whether the platform's algorithm exhibits differential behavior in response to geopolitical versus everyday content.

3.1 Keyword Selection

The South China Sea (SCS) dataset was curated through expert consultations with researchers from the Atlantic Council's Digital Forensic Research Lab, the National University of Singapore, and De La Salle University. Keywords focus on regional disputes, military activity, and environmental issues [2, 7]. For the General Content dataset, we followed YouTube's content taxonomy [13], covering a wide spectrum from entertainment and sports to science and education. Table 1 summarizes the keyword sets.

Table 1: Keywords used for data collection. Representative keywords are shown for the South China Sea; full categories are listed for General Content.

Dataset	Keywords
South China Sea	South China Sea dispute, China nine dash line, China artificial islands, US Navy South China Sea, China maritime claims, South China Sea latest news, China vs Philippines, South China Sea oil and gas, ASEAN South China Sea talks, China territorial dispute.
General Content	Film & Animation, Autos & Vehicles, Music, Pets & Animals, Sports, Travel & Events, Gaming, People & Blogs, Comedy, Entertainment, News & Politics, How-to & Style, Education, Science & Technology, Nonprofits & Activism.

3.2 YouTube Shorts Collection

Since the YouTube Data API v3 does not support Shorts, we used APIFY's YouTube Scraper [34] to collect Shorts video IDs based on the keywords in Table 1. The scraper was configured to retrieve only Shorts by enabling the `maxResultsShorts` parameter and disabling other video types. We applied no date filters and used the `relevancy` sort option to prioritize videos most relevant to the keywords. We collected 500 seed videos for each domain: South China Sea and General Content.

3.3 Recommendation Collection

To collect Shorts recommendations, we developed a custom scraping framework, as existing tools do not support this functionality. Each seed video was opened in a fresh Selenium-driven browser session with no login, cookies, or browsing history to simulate a neutral user environment. We simulated user interaction by scrolling through recommended Shorts up to a depth of 10, where depth refers to the position in the recommendation chain. The browser was fully reset after each session to avoid cross-session contamination. From each dataset, we collected 5,000 recommended videos. Combined with the 500 seed videos, each dataset contains 5,500 Shorts, totaling 11,000 videos overall.

3.4 Keyframe Collection

To extract representative visuals from each YouTube Short, we employed the PRISM framework [5], which identifies keyframes based on perceptual changes aligned with human visual sensitivity. The framework takes a video as input and returns visually salient frames that represent important moments. This perceptually guided approach is well-suited for Shorts content, where visual impact and rapid engagement are central to viewer experience. By focusing on human-like perception, PRISM helps surface frames that are more likely to capture user attention and potentially influence recommendation dynamics, especially in politically sensitive or emotionally charged videos. An example of extracted keyframes is shown in Figure 1.

We selected PRISM for its strong performance across accuracy, fidelity, and compression efficiency, as summarized in Table 2, making it a practical and scalable choice for high-volume Shorts analysis.

Figure 1: Keyframes extracted using the PRISM framework.

Table 2: Performance comparison of keyframe extraction models with reported accuracy, fidelity, and compression ratio (CR).

Model	Accuracy (%)	Fidelity (%)	CR (%)
PRISM [5]	85.58	70.30	99.23
LiveLight [38]	72.30	80.00	90.00
DSVS [10]	66.00	75.00	95.00

The final number of videos and corresponding keyframes used in this study are summarized in Table 3. These keyframes form the basis for downstream analysis of algorithmic behavior and content characteristics.

Table 3: Summary of videos, frames, and extracted keyframes per dataset.

Dataset	Total Videos	Frames	Keyframes
South China Sea	5,500	338,246	15,361
General Content	5,500	326,184	14,592
Total	11,000	664,430	29,953

4 Methodology

In this section, we describe our approach for generating captions from keyframes and embedding both modalities to analyze YouTube Shorts recommendations.

4.1 Keyframe Caption Generation

To support multimodal analysis, we generate natural language captions for each keyframe, providing a complementary semantic signal to the visual content. This aids interpretability and clustering, especially given the dense and fast-paced nature of Shorts.

We use the `Llama-3.2-11B-Vision-Instruct` [31] model, a multimodal, instruction-tuned system designed for tasks like image captioning. It performs competitively in zero-shot settings on benchmarks such as VQAv2 and TextVQA (Table 4), making it well-suited for producing high-quality captions for short-form content.

Table 4: Validation performance of multimodal models on VQAv2 and TextVQA. VQAv2 tests general visual QA, while TextVQA focuses on understanding text within images.

Model	VQAv2 [15]	TextVQA [33]
LLaMA 3.2 11B [31]	66.8	73.1
Flamingo-80B [1]	56.3	35.0
IDEFICS-9B [24]	50.9	25.9

4.2 Multimodal Embedding with CLIP

To convert both keyframes and their generated captions into a unified semantic representation, we used the OpenCLIP implementation of the CLIP ViT-G/14 model [35]. This model was selected for its state-of-the-art zero-shot performance on vision-language benchmarks (Table 5). Keyframe images were embedded using CLIP's vision encoder, while the generated captions were embedded using its text encoder. All embeddings were L2-normalized to enable alignment in a shared semantic space. This facilitates downstream tasks such as clustering, where both visual and textual signals contribute to identifying algorithmic patterns. Unlike supervised models trained on fixed label sets, CLIP's zero-shot approach is more generalizable to open-ended and dynamic content domains like YouTube Shorts, where new trends and topics emerge rapidly.

Table 5: Zero-shot benchmarks for vision-language models. Metrics: ImageNet Top-1 (ZS), COCO R@5 (ZS), Flickr30k R@5 (ZS).

Model	#Params	ImageNet [12] Top-1 (ZS,%)	COCO [27] R@5 (ZS,%)	Flickr30k [36] R@5 (ZS,%)
CLIP ViT-G/14 [32, 35]	1.8B	80.2	75.0	78.5
OpenCLIP ViT-H/14 [9]	1.2B	78.0	73.4	75.8
OpenCLIP RN50×64 [9]	435M	70.4	62.5	63.9
CLIP ViT-L/14 [32]	427M	75.3	70.2	72.5
CLIP ViT-B/32 [32]	150M	63.3	60.5	62.0
ALIGN [20]	1.8B	76.4	77.0	79.2
BLIP [26]	213M	74.3	72.4	74.7
SigLIP ViT-B/16 [40]	86M	74.0	68.5	70.0

All metrics are reported in zero-shot settings.

5 Results

We visualize the structure of keyframe and caption embeddings using UMAP [30], a non-linear dimensionality reduction method that preserves meaningful relationships in high-dimensional data. This enables intuitive comparisons between seed and recommended content across domains and modalities.

Figure 2 shows UMAP projections for the politically sensitive *South China Sea* (SCS) dataset and general YouTube content. In the SCS domain (Figures 2a and 2b), seed videos (blue points and hulls) form a tight cluster, indicating a focused topical anchor. In contrast, recommended videos (red points) are more dispersed and frequently fall outside the seed region, suggesting that YouTube's algorithm introduces content that diverges in theme, style, or presentation. This supports **RQ1**, confirming that recommendations shift away from their seed content, particularly in politically sensitive contexts.

For general YouTube content (Figures 2c and 2d), the seed distribution is broader, yet recommendations remain closer and more overlapping, indicating a more stable and consistent recommendation pattern. This aligns with **RQ3**, showing that the degree of shift varies by domain: sensitive topics experience greater divergence, while general content remains more coherent.

Keyframe and caption embeddings exhibit parallel patterns, with independently derived clusters reinforcing one another. This supports **RQ2**, demonstrating that keyframes and their captions together provide interpretable signals of recommendation behavior and potential bias.

(a) **Visual embeddings of keyframes (SCS)**

(b) **Caption embeddings of keyframes (SCS)**

(c) **Visual embeddings of keyframes (General)**

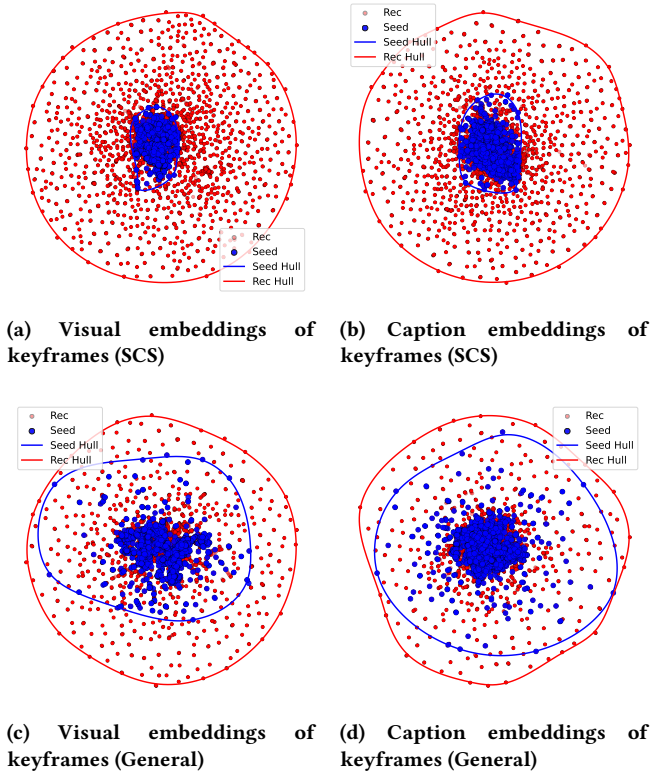(d) **Caption embeddings of keyframes (General)**

Figure 2: UMAP projections of keyframe and caption embeddings for South China Sea and general YouTube content. Each pair compares the distribution between seed and recommended videos.

To support the visual analysis, we compute quantitative metrics that capture structural differences between seed and recommended video embeddings. Table 6 shows that in the general dataset, recommendation variance increases moderately (e.g., 46.91 for captions and 59.69 for frames). In contrast, the SCS dataset exhibits substantially higher spread, with variances of 79.13 (captions) and 89.13 (frames), as well as greater intra-cluster distances (15.63 and 16.74), indicating more dispersed and thematically inconsistent recommendations. These differences suggest that recommendations originating from politically sensitive content tend to drift more broadly in both visual and textual dimensions.

Table 7 quantifies the divergence between the South China Sea and general YouTube datasets. Using Jensen-Shannon Divergence and Wasserstein Distance to capture global distributional shifts, and Δ metrics to assess structural spread, we observe consistently higher scores in frame space (e.g., JSD = 0.889, Δ variance = 0.909). This suggests that visual features diverge more sharply than textual ones across the two domains. While caption embeddings also show notable divergence, the differences are less pronounced than in the visual space. One interpretation is that visual framing, such as imagery, style, or scene composition, varies more between sensitive and general content than caption-level semantics. This further underscores the value of analyzing both modalities and highlights how visual presentation may subtly reflect topic sensitivity in algorithmic recommendations.

Table 6: Normalized divergence scores comparing South China Sea and general YouTube datasets across embedding spaces. Frame embeddings show stronger divergence than captions.

Metric	General		South China Sea (SCS)	
	Caption	**Frame**	**Caption**	**Frame**
Seed Variance	7.82	12.86	4.29	4.23
Rec Variance	46.91	59.69	79.13	89.13
Seed Intra Dist	4.70	6.11	3.50	3.48
Rec Intra Dist	11.55	13.32	15.63	16.74

Table 7: Normalized divergence scores for SCS and General datasets. Lower values indicate minimal difference; higher values reflect stronger divergence between seed and recommendation distributions. JSD and Wasserstein capture global shifts, while Δ metrics reflect clustering changes.

Metric	Caption Space	Frame Space
Jensen-Shannon Divergence	0.708	0.889
Wasserstein Distance (scaled)	0.552	0.602
Normalized Δ Variance	0.662	0.909
Normalized Δ Intra-Dist	0.629	0.870

Together, these visual and quantitative results offer a consistent, multimodal perspective on how recommendation behavior varies across domains. Our keyframe-based approach captures these shifts effectively and presents a scalable, interpretable method for auditing content dynamics in short-form video platforms.

6 Conclusion

In this study, we introduced a keyframe-based framework for analyzing YouTube Shorts recommendation behavior using both visual and textual signals. By extracting perceptually salient frames and generating semantic embeddings, we traced how content shifts through the recommendation pipeline. Our findings reveal substantial content drift in politically sensitive topics such as the South

China Sea, in contrast to the more stable recommendation patterns observed in general content. The key contribution of this work is demonstrating that keyframes serve as a scalable and interpretable proxy for detecting algorithmic bias without requiring full video processing. While this study focuses on two domains, future research could broaden the scope to include additional topics and incorporate complementary analyses such as topic modeling, sentiment tracking, or engagement signals. Overall, our results underscore how recommender systems can subtly shape content exposure and reinforce the importance of transparency in algorithmic media distribution.

GenAI Usage Disclosure

The authors used generative AI tools as part of the research methodology. Specifically, the LLaMA-3.2-11B-Vision-Instruct model was used to generate textual captions from video keyframes for content analysis. No generative AI tools were used in the writing, coding, or data collection processes beyond this purpose.

Acknowledgments

References

[1] Jean-Baptiste Alayrac et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35, 23716–23736.

[2] Atlantic Council. 2025. Digital forensic research lab. https://www.atlanticcouncil.org/programs/digital-forensic-research-lab/. Accessed: 2025-02-25. (2025).

[3] Mert Can Cakmak and Nitin Agarwal. 2025. Unpacking algorithmic bias in youtube shorts by analyzing thumbnails. In *Proceedings of the 58th Hawaii International Conference on System Sciences*. (Jan. 2025). https://hdl.handle.net/10125/109144.

[4] Mert Can Cakmak, Nitin Agarwal, and Remi Oni. 2024. The bias beneath: analyzing drift in youtube's algorithmic recommendations. *Social Network Analysis and Mining*, 14, 1, 171.

[5] Mert Can Cakmak, Nitin Agarwal, and Diwash Poudel. 2025. Prism: perceptual recognition for identifying standout moments in human-centric keyframe extraction. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. Accepted for presentation. Copenhagen, Denmark.

[6] Mert Can Cakmak, Obianuju Okeke, Ugochukwu Onyepunuka, Billy Spann, and Nitin Agarwal. 2023. Investigating bias in youtube recommendations: emotion, morality, and network dynamics in china-uyghur content. In *International Conference on Complex Networks and Their Applications*. Springer, 351–362.

[7] Centre for International Law, National University of Singapore. 2025. Ocean law and policy – south china sea. https://cil.nus.edu.sg/research/ocean-law-policy/south-china-sea/. Accessed: 2025-02-25. (2025).

[8] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: a survey and future directions. *ACM Transactions on Information Systems*, 41, 3, 1–39.

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Also see the open_clip codebase at https://github.com/mlfoundations/open_clip, 2818–2829.

[10] Yang Cong, Junsong Yuan, and Jiebo Luo. 2011. Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Transactions on Multimedia*, 14, 1, 66–75.

[11] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, 191–198.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: a large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[13] EntreResource. 2023. Youtube video categories – full list explained and which you should use in 2023. https://entreresource.com/youtube-video-categories-full-list-explained-and-which-you-should-use/. Accessed: 2025-02-25. (2023).

[14] Rajesh Garapati and Manomita Chakraborty. 2025. Recommender systems in the digital age: a comprehensive review of methods, challenges, and applications. *Knowledge and Information Systems*, 1–45.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

[16] Georg Groh, Stefan Birnkammerer, and Valeria Köllhofer. 2012. Social recommender systems. *Recommender systems for the social web*, 3–42.

[17] Hussam Habib and Rishab Nithyanand. 2025. Youtube recommendations reinforce negative emotions: auditing algorithmic bias with emotionally-agentic sock puppets. *arXiv preprint arXiv:2501.15048*.

[18] Muhammad Haroon, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, Zubair Shafiq, and Magdalena Wojcieszak. 2022. Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations. *arXiv preprint arXiv:2203.10666*.

[19] Hazem Ibrahim, HyunSeok Daniel Jang, Nouar Aldahoul, Aaron R Kaufman, Talal Rahwan, and Yasir Zaki. 2025. Tiktok's recommendations skewed towards republican content during the 2024 us presidential race. *arXiv preprint arXiv:2501.17831*.

[20] Chao Jia et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[21] Baris Kirdemir and Nitin Agarwal. 2021. Exploring bias and information bubbles in youtube's video recommendation networks. In *International Conference on Complex Networks and Their Applications*. Springer, 166–177.

[22] Baris Kirdemir, Joseph Kready, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. 2021. Examining video recommendation bias on youtube. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 106–116.

[23] Baris Kirdemir, Joseph Kready, Esther Mead, Muhammad Nihal Hussain, Nitin Agarwal, and Donald Adjeroh. 2021. Assessing bias in youtube's video recommendation algorithm in a cross-lingual and cross-topical context. In *Social, Cultural, and Behavioral Modeling: 14th International Conference, SBP-BRiMS 2021, Virtual Event, July 6–9, 2021, Proceedings 14*. Springer, 71–80.

[24] Hugo Laurençon et al. 2023. Introducing idefics: an open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics. Hugging Face. (2023).

[25] Mark Ledwich and Anna Zaitsev. 2019. Algorithmic extremism: examining youtube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 740–755.

[28] Xiao Lin, Xiaokai Chen, Linfeng Song, Jingwei Liu, Biao Li, and Peng Jiang. 2023. Tree based progressive regression model for watch-time prediction in short-video recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4497–4506.

[29] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 2145–2148.

[30] Leland McInnes, John Healy, and James Melville. 2018. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[31] Meta. 2025. Meta-llama/llama-3.2-11b-vision-instruct. Hugging Face. Accessed: January 4, 2025. (2025). https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.

[32] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[33] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

[34] Streamers. 2024. Youtube scraper. Accessed: 2024-01-10. APIFY. https://apify.com/streamers/youtube-scraper.

[35] Mitchell Wortsman. 2023. Reaching 80% zero-shot accuracy with openclip: vit-g/14. https://laion.ai/blog/giant-openclip/. (2023).

[36] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2, 67–78.

[37] Ruohan Zhan, Changhua Pei, Qiang Su, Jianfeng Wen, Xueliang Wang, Guanyu Mu, Dong Zheng, Peng Jiang, and Kun Gai. 2022. Deconfounding duration bias in watch-time prediction for video recommendation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 4472–4481.

[38] Bin Zhao and Eric P Xing. 2014. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2513–2520.

[39] Yu Zheng, Chen Gao, Jingtao Ding, Lingling Yi, Depeng Jin, Yong Li, and Meng Wang. 2022. Dvr: micro-video recommendation optimizing watch-time-gain under duration bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, 334–345.

[40] Wengang Zhou, Weichao Zhao, Hezhen Hu, Zecheng Li, and Houqiang Li. 2024. Scaling up multimodal pre-training for sign language understanding. *arXiv preprint arXiv:2408.08544*.